IBM AI Ethics Board

Foundation models: Opportunities, risks and mitigations

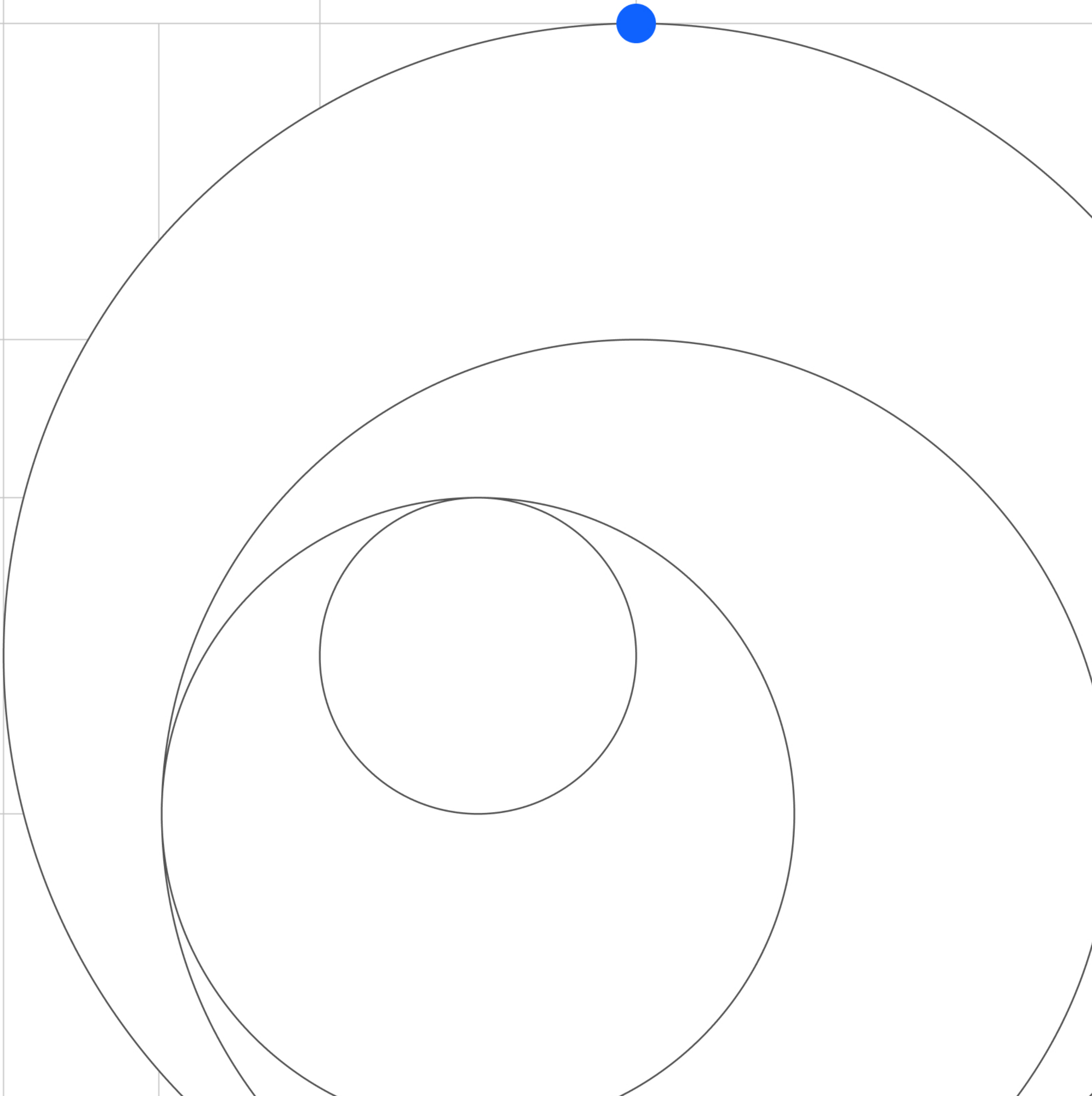


Table of contents

03 Executive summary

04 Introduction

05 Benefits of foundation models

07 Risks of foundation models

11 Principles, pillars and governance

12 Guardrails and mitigations

14 AI policies, regulation and best practices

Executive summary

The rise of foundation models offers enterprises exciting new possibilities but also raises new and expanded questions about their ethical design, development, deployment and use. According to a recent IBM Institute for Business Value generative AI survey, organizations are already expressing concerns about trust-related issues—specifically as barriers to investment. Their top concerns are cybersecurity (57%), privacy (51%) and accuracy (47%). Many organizations were taking these concerns seriously before the *consumerization* of generative AI, expressing their intent to invest at least 40% more in AI ethics over the next three years. Awareness about risks and possible ways to mitigate them is the first crucial step toward building trustworthy AI systems.

In this document we:

Explore the benefits of foundation models, including their capability to perform challenging tasks, potential to speed up the adoption of AI, ability to increase productivity and the cost benefits they provide.

Discuss the three categories of risk, including risks known from earlier forms of AI, known risks amplified by foundation models and emerging risks intrinsic to the generative capabilities of foundation models.

Cover the principles, pillars and governance that form the foundation of IBM's AI ethics initiatives and suggest guardrails for risk mitigation.

Introduction

As the use of AI continues to expand, large and complex AI models are delivering promising performance results, as well as solving some of society's most challenging problems. However, building large training data sets and complex models for each AI application can be burdensome for enterprises. Foundation models provide a path to achieve the best of both worlds: build powerful state-of-the-art models and reuse them directly or apply tuning methods to implement a variety of use cases, rather than train new models for each use case. For example, IBM Research® developed foundation models for visual inspection. These foundation models learn the general representation of concrete surfaces and runways and can be further tuned for specific use cases like crack detection or defect inspection with less labeled data.

IBM defines a *foundation model* as an AI model that can be adapted to a wide range of downstream tasks. Foundation models are typically large-scale generative models that are trained on unlabeled data using self-supervision. As large-scale models, foundation models can include billions of parameters.

IBM is a hybrid cloud and AI company with a long reputation as a responsible data steward committed to AI ethics. Using the strength of our research, product and consulting teams, along with external partners, such as Hugging Face, we help bring the power of foundation models to our clients and build trustworthy AI across any enterprise. IBM also continues to invest in building new platforms, such as the IBM® watsonx™ AI and data platform and technologies, for designing and developing AI models to behave in an auditable and trustworthy manner.

This document describes the point of view of IBM on the ethics of foundation models. It is the first version, and future versions will expand on various aspects of IBM's foundation model ethics approach. We hope this document is helpful for all stakeholders in developing, deploying and using the foundation model in a responsible way.

Benefits of foundation models

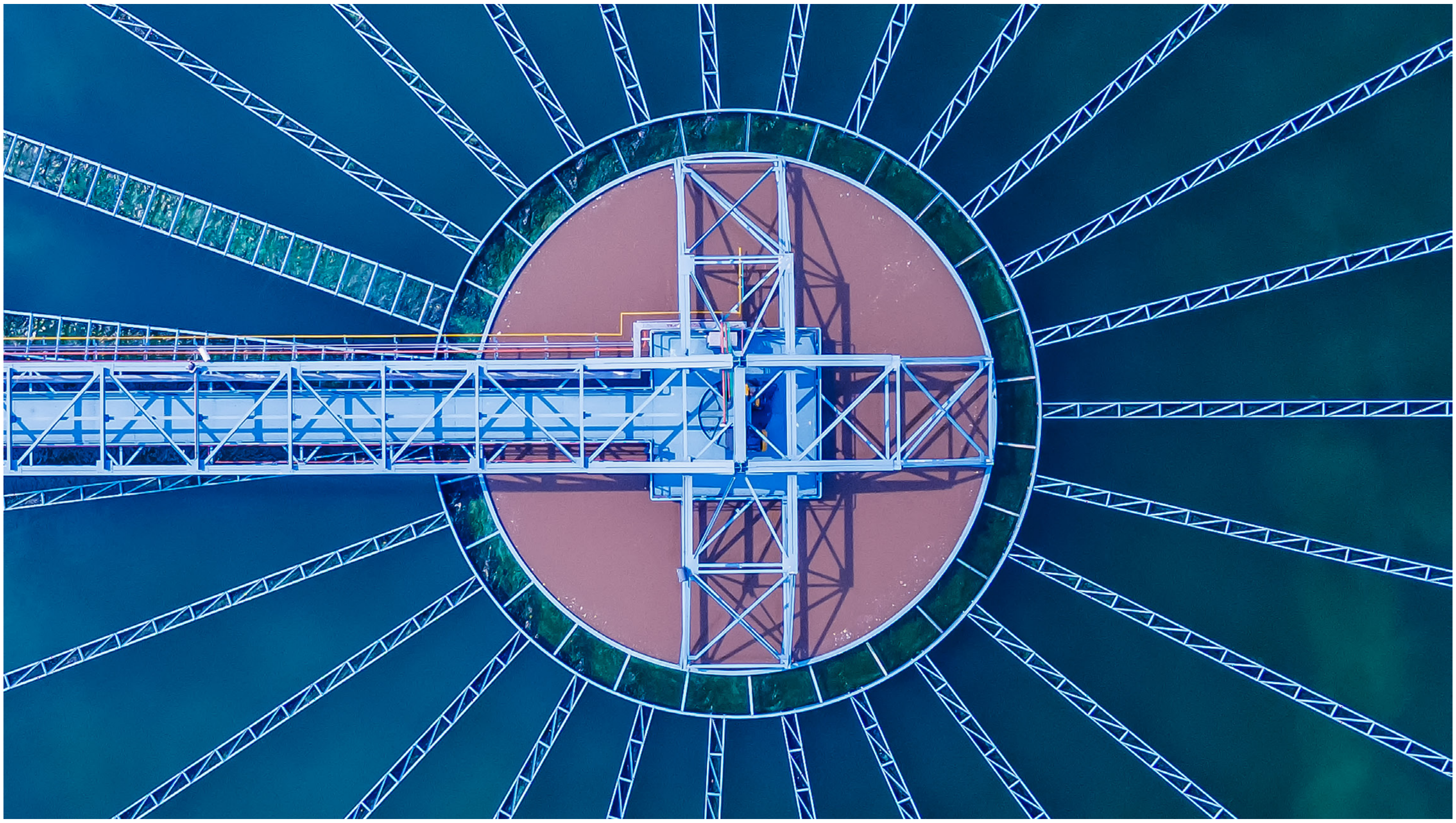
Foundation models can significantly improve the process of developing AI systems and help advance AI from the exploration to the adoption phase in enterprises. Their benefits include:

Performing complex tasks
Foundation models show a significant increase in performance in solving difficult and complex problems. For example, the geospatial foundation model from the IBM and NASA collaboration is designed to convert NASA's satellite data into maps of natural disasters like floods and other landscape changes. The model could also be used to help reveal our planet's past; estimate risks to crops, businesses or infrastructures due to severe weather; develop strategies to adapt to climate change; and assist agribusiness. The model is planned to be made available in preview to IBM clients through the IBM Environmental Intelligence Suite.

As another example, IBM's MoLFormer-XL is a foundation model that infers the structure of molecules from simple representations and makes it easy to learn various downstream tasks like predicting a molecule's physical and quantum properties, identifying similar molecules, screening already approved molecules for new use cases, and discovering new molecules. Moderna and IBM are exploring ways to use MoLFormer to help predict molecule properties and understand the characteristics of potential mRNA medicines.

Increased productivity
The generative nature of foundation models expands the number of areas where AI can be used in an enterprise to help improve productivity by automating routine and tedious tasks and allowing users to spend more time on creative and innovative work. For example, IBM Watson® Code Assistant, powered by foundation models, enables developers of all experience levels to write code using AI-generated recommendations.

Quicker time to value
Foundation models are usually trained with unlabeled data, which is more accessible in larger quantities than labeled data. Once trained, foundation models can be used either directly or after being tuned for downstream applications, using a small amount of specialized labeled data, which can decrease time-to-value creation.



Utilize diverse data modalities
Foundation models may be trained using various data modalities, such as natural language, text, image and audio. They can also be applied to tasks requiring different data types, such as time series data, geospatial data, tabular data, semi-structured data and mixed-modality data like text combined with images.

Amortized expenses
Although the initial cost of training a foundation model is significantly higher than training a traditional AI model, the incremental cost of applying it to a new task is considerably lower. Using pretrained foundation models could help eliminate the requirement that enterprises make substantial investments to train foundation models to experiment with their new capabilities. For an enterprise, the trustworthiness of the models, energy efficiency, performance, portability and the ability to use enterprise data effectively and securely are paramount.

IBM allows enterprises to create and own the value of foundation models for their business by bringing the best innovations from the open, global AI community, running efficiently in hybrid computing environments, helping mitigate risks, and rigorously governing AI.

Risks of foundation models

Like all rapidly advancing technologies, foundation models have risks along with benefits. Some are legal risks, for example, restrictions on moving or using data, and need to be carefully evaluated under current and evolving law. Other risks have an ethical nature and must be considered carefully so that the technology has a positive impact. In general, AI risks raise sociotechnical questions and should be addressed and mitigated through sociotechnical methods, including software tools, risk assessment processes, AI ethics frameworks, governance mechanisms, multistakeholder consultations, standards and regulation. We can easily demonstrate the risks by considering the following 3 categories:

1. **Traditional.** Known risks from prior or earlier forms of AI systems
2. **Amplified.** Known risks but now intensified because of intrinsic characteristics of foundation models, most notably their inherent generative capabilities
3. **New.** Emerging risks intrinsic to foundation models and their inherent generative capabilities

This information is useful, since there already are mitigation mechanisms that could be helpful in partially addressing the traditional and amplified risks. We also structure the list of risks in relation to whether they're mostly associated with content provided to the foundation model—the input—or the content generated by it—the output—or if they're related to additional challenges.

1. Risks associated with input

	Group	Risk	Indicator
Training and tuning phase	Fairness	Bias like historical, representational or measurement bias	Amplified
	Robustness	An adversary or malicious insider injecting false, misleading, malicious or incorrect samples	Traditional
	Value alignment	Using undesirable output, such as inaccurate or inappropriate user content, from downstream applications for retraining purposes	New
	Data laws	Legal restrictions on moving or using data	Traditional
	Intellectual property	Copyright and other IP issues with the content	Amplified
	Transparency	The ability to disclose what content is collected, how it will be used and stored, and who has access	Amplified
	Privacy	Inclusion or presence of personal identifiable information and sensitive personal information	Traditional
		Challenges around the ability to provide data subject rights, for example, opt out, right to access or right to be forgotten	Amplified
Inference phase	Privacy	Disclosing personal information or sensitive personal information as a part of prompt sent to the model	New
	Intellectual property	Disclosing copyright information or other IP information as part of the prompt sent to the model	New
	Robustness	Vulnerabilities to adversarial attacks like evasion, which is an attempt to make a model output incorrect by perturbing the data sent to the trained model	Amplified
		Vulnerabilities to adversarial attacks like prompt injection, which forces a different output to be produced; prompt leaking, which is the disclosure of the system prompt; or jailbreaking, which is avoiding guardrails established in the system prompt	New

2. Risks associated with output

Group	Risk	Indicator
Fairness	Bias in the generated content	New
	Performance disparity across individuals or groups	Traditional
Intellectual property	Copyright infringement, including compliance with open-source license agreements	New
Value alignment	Hallucination—false content generation	New
	Toxic, hateful, abusive and aggressive output	New
Misuse	Spread disinformation—deliberate creation of misleading information	Amplified
	Generate toxic, hateful, abusive and aggressive content	New
	Nonconsensual use of people's likeness—deepfakes	Amplified
	Dangerous use—creating plans to develop weapons or malware	New
	Deceptive use of generated content—intentional nondisclosure of AI-generated content	New
Harmful code generation	Execution of harmful generated code	New
Privacy	Exposing personal information or sensitive personal information in generated content	New
Explainability	Challenges in explaining why output was generated	Amplified
Traceability	Challenges in determining original source and facts of the generated output	New

3. Challenges

	Group	Risk	Indicator
Governance	Transparency	Documenting data and model details, purpose, potential uses and harms	Traditional
	Accountability	Determining responsibility for misaligned generated output along the AI lifecycle and value chain	Amplified
Legal compliance	Intellectual property	Determining creator of downstream models	New
		Determining creator of open-source foundation models	New
		Determining ownership of AI-generated content	New
		Legal uncertainty about intellectual property rights related to generated content	New
	Legal uncertainty	Determining downstream obligations	Amplified
Societal impact	Impact on jobs	Human displacement—AI-induced job loss—and reskilling, which can particularly impact creative jobs or jobs with repetitive tasks	Amplified
	Human dignity	Human exploitation—ghost work in training; inadequate working conditions; lack of healthcare, including mental health; unfair compensation	Amplified
	Environment	Increased carbon emission due to high energy requirement to train and operate	Amplified
	Diversity and inclusion	Homogenizing of culture and thoughts	New
	Human agency	Misinformation and disinformation generated by foundation models, including generation of manipulative content	Amplified
	Impact on education	Bypassing the learning process and plagiarism	New

Principles, pillars and governance

IBM's Principles for Trust and Transparency and Pillars for trustworthy AI are the foundation for IBM's AI ethics initiatives. IBM has an AI Ethics Board with the mission to support a centralized governance, review and decision-making process for IBM AI ethics policies, practices, communications, research, products and services. The board includes a diverse set of stakeholders from across the company and is supported by a community of IBM employees who serve as AI focal points and AI ethics advocates. Through the board, IBM's principles are put into practice. As new technology emerges, such as foundation models, the IBM AI Ethics Board is actively engaged in supporting alignment with these Principles and Pillars, which evolve to address new AI ethics issues.

Guardrails and mitigations

IBM has established an organizational culture that supports the responsible development and use of AI. Based on the IBM Institute for Business Value AI ethics in action report, AI ethics has already become more business-led versus technology-led, and nontechnical executives are now the primary champions for AI ethics, increasing from 15% in 2018 to 80% 3 years later. Additionally, 79% of CEOs are now prepared to act on AI ethics issues, up from 20%. We recognize that responsible AI is a sociotechnical area that requires a holistic investment in culture, processes and tools. Our investment in our own organizational culture includes assembling inclusive, multidisciplinary teams and establishing processes and frameworks to assess risks.

IBM is engaging in cutting-edge research and developing tools to help support professionals throughout the lifecycle of responsible and trustworthy AI. The watsonx enterprise-ready AI and data platform is built with 3 components: the IBM watsonx.ai™ AI studio, IBM watsonx.data™ data store and IBM watsonx.governance™ toolkit. IBM's AI governance technology enables users to drive responsible, transparent and explainable AI workflows. This technology includes IBM Watson OpenScale, which tracks and measures outcomes from AI models through their lifecycle and helps organizations monitor fairness, explainability, resiliency, alignment with business outcome and compliance. IBM has also developed several methods to help with bias issues like FairIJ, Equi-tuning, and FairReprogram. Read more about additional open-source trustworthy AI tools.

Additional guardrails and mitigations include:

Transparency reporting
Using standardized factsheet templates is one way to accurately log details of the data and model, purpose, and potential use and harms.
Read more here →

Filtering undesirable data
Using curated, higher-quality data can help mitigate certain issues. IBM is developing filtering techniques to help reduce the chances of producing undesirable, misaligned content by removing hate language, biased language and profanity from the data.
Read more here →

Domain adaptation
Training a foundation model to a specific domain or industry can help minimize the scope of risk the models can give rise to because it can be conditioned to generate outputs that are tuned to be more relevant to that domain or industry.
Read more here →

Human oversight and human in the loop
Human oversight and review can help identify and correct errors and biases in the generated output. Also, human validation and feedback on the quality of model responses help ensure that the generated content is accurate, relevant, of high quality, not drifting and aligned.
Read more here →

Consulting engagement
IBM Consulting™ is dedicated to helping clients with the safe and responsible use of AI irrespective of the preferred tech stack. They help clients nurture a culture that adopts and scales AI safely, creates investigative tools to see inside black box algorithms and makes sure clients' corporate strategy includes strong data governance principles.
Read more here →

IBM Enterprise Design Thinking
IBM Enterprise Design Thinking® methods and frameworks, such as Team Essentials for AI, help clients define ethical behaviors throughout the AI design and development process.
Read more here →

AI Ethics review
Assessment of capabilities, limitations and risks in AI projects helps ensure the responsible development and use of the technology.

Ethics by Design
Ethics by Design is a structured framework with the goal of integrating tech ethics in the technology development pipeline, including, but not limited to, AI systems. Ethics by Design enables AI and other technologies as a force for good by embedding tech ethics principles throughout products, services and broader operations.

Team diversity
Diversity in the teams that build and train AI systems, including foundation models, helps ensure that a variety of perspectives and experiences are considered. This diversity improves the accuracy and performance of AI systems and helps reduce risks throughout the AI lifecycle, including the potential for adverse outcomes that impact groups that may not be well represented on less diverse teams.

AI policies, regulation and best practices

A Policymaker's Guide to Foundation Models introduces what policymakers need to know about foundation models. This blog, from the IBM Policy Lab, aims to help policymakers in the complex task of regulating the use of generative AI, aiming to avoid the risks without limiting innovation and beneficial opportunities. For additional information on IBM's recommendations to policymakers, read IBM Chief Privacy and Trust Officer Christina Montgomery's testimony before the U.S. Senate Judiciary Subcommittee on Privacy, Technology and the Law here.

IBM is making an impact in shaping regulatory policy, industry best practices and tools, governance of emerging technologies, and sociotechnical research by leading and contributing to initiatives with organizations, such as:

- The World Economic Forum
- Partnership on AI
- The International Association of Privacy Professionals (IAPP) AI Governance Center
- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- Christina Montgomery's service on the National Artificial Intelligence Advisory Committee (NAIAC)
- The United Nations Global Digital Compact
- The Global Partnership on Artificial Intelligence (GPAI)
- The Organisation for Economic Co-operation and Development (OECD)
- The Data & Trust Alliance

IBM has strong academic partnerships like the MIT-IBM Watson AI Lab, where a community of scientists at MIT and IBM Research conducts AI research and works with global organizations to bridge algorithms to their impact on business and society. The Notre Dame-IBM Tech Ethics Lab was formed to address the many diverse ethical questions implicated by the development and use of advanced technologies, including AI, machine learning (ML) and quantum computing. The Stanford University Human-Centered Artificial Intelligence (HAI) research advances AI research, education, policy and practices.

Keep watching this space to learn more about the latest developments in foundation models and how IBM is working toward the responsible development and use of this and other technologies.

© Copyright IBM Corporation 2023

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
August 2023

IBM, the IBM logo, Enterprise Design Thinking, IBM Consulting, IBM Research, IBM Watson, watsonx, watsonx.ai, watsonx.data, and watsonx.governance are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is compliant with any law or regulation. Statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.